DOCUMENT RESUME

ED 189 969 HE 012 892

AUTHOR Weber, Margaret B.
TITLE Implications of Truth-in-Testing Legislation on Student Selection. AIR Forum 1980 Paper.
PUB DATE Apr 80
NOTE 29p.: Paper presented at the Annual Forum of the Association for Institutional Research (20th, Atlanta, GA, April 27-May 1, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Admission Criteria: College Admission: *College Entrance Examinations: Comparative Testing: *Disclosure: Equated Scores: Higher Education: Information Dissemination: Legal Problems: *Scoring: Scoring Formulas: *Standardized Tests: State Legislation: Student Evaluation
IDENTIFIERS *AIR Forum 1980: *Truth in Testing

ABSTRACT

Some of the technical issues associated with legislation to regulate testing and one alternative to test scores in the admissions and student selection process are discussed. Technical issues that must be addressed in order to accommodate full and public disclosure include equating, test administration, scoring, and score reporting. The need to equate tests is a consequence of the need to have multiple versions of forms of the same test. Equating is the process that allows comparing scores from one form of a test to scores obtained from another form of the test. The following methods of test equating used in large-scale testing programs are described: spiraling, anchor-test equating, pre-section equating, and item-response theory equating. The problem faced by test developers of meeting disclosure requirements without making tests prohibitively expensive is considered. Test administration for the handicapped and adjustments required for special administration (e.g., for religious reasons) are noted. The meaning of rights scoring and formula scoring is explained, and the legislative requirement concerning reporting to examinees is viewed as an opportunity to educate the public about measurement. An alternative to test scores for student selection is a standardized structured screening procedure conducted by trained examiners. (SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED189969

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

IMPLICATIONS OF TRUTH-IN-TESTING LEGISLATION ON STUDENT SELECTION

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

AIR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Margaret B. Weber

Southern Regional Office
Educational Testing Service
3445 Peachtree Road, N.E.
Suite 1040
Atlanta, Georgia 30326

Paper prepared for the
Association for Institutional
Research Annual Meeting
Atlanta, Georgia
April 28-30, 1980

HE 012 892

*Parts of this paper were presented at the Southern Regional Assembly of the College Board in Biloxi, Mississippi, February 19-21, 1980.

THE ASSOCIATION FOR INSTITUTIONAL RESEARCH

This paper was presented at the Twentieth Annual Forum of the Association for Institutional Research held at the Peachtree Plaza Hotel in Atlanta, Georgia, April 27 - May 1, 1980. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of Forum papers.

Mary Corcoran
University of Minnesota
(Editor, AIR Forum Publications)

Even to a casual observer, two important outcomes are evident from the current legislative activity in the area of testing. First, consumers of tests are much more aware of and concerned about testing issues than they used to be. Second, testing professionals have greater incentive to give serious consideration to longstanding psychometric practice. These outcomes, not only desirable individually but also complementary, should result in a dialogue between consumers and producers of tests that will enhance the use and development of tests. Many of the new testing laws will require some change either in how tests are developed or in how and when test developers report to the public. Under these circumstances, a discussion of technical issues that may be affected by testing legislation seems especially pertinent.

To participants in the business of admissions and student selection a third outcome must be discussed. That is, what are efficient, reliable, and valid alternatives to test scores in the admissions process. The purpose of this paper is to discuss some of the technical issues associated with legislation to regulate testing and to discuss one alternative to test scores in the admissions and student selection process.

Technical Issues

As an example of legislation which has only a moderate impact on testing practice, the Dunlap bill (California S.B . 2005) requires that copies of unsecure tests or test samples equivalent to secure tests in current use be made a matter of public record. This requirement may not necessitate changes in traditional test development procedures, administration, scoring or equating. It may, however, require changes in reporting practices. For some testing programs, compliance with the California bill might result in having sample tests, already available to test candidates, introduced as part of the public record. For other testing programs,

compliance might necessitate the production of an extra test form, not for administration but for release as public information. And, for still other testing programs, a previously secure, intact test form might have to be released for the public record, thereby reducing the number of secure test forms available for administration. As these illustrations indicate, the California bill is likely to affect the cost or number of administrations but would not directly require departures from the traditional practice of test development.

Other legislation includes more stringent requirements intended to regulate test professionals. In particular, the issue of disclosure of all tests after use will prompt test professionals to consider changes in traditional practice. This requirement stipulates that the questions and answers for a given test form must be released to the candidate upon request, usually within thirty days of the score release date. It is this requirement for <u>full and public disclosure</u> found first in the unsuccessful Maryland House Bill 1425, that is likely to bring considerable change to test development practice in the decade of the eighties.

It is important that discussions of the full and public disclosure requirement not be resticted to test developers. It is imperative that these discussions occur among the publics who take or use tests. It is imperative that these "consumers" interact with the test developers. Indeed, it is much more significant -- to take liberty with a technical term -- that the publics who are affected by tests are now actively interested in <u>how</u> and <u>why</u> tests are developed, <u>how</u> they are appropriately <u>used</u>, and <u>what</u> their limitations are than it is that there is now legislated regulation of testing.

This increased vocal public interest gives test developers an opportunity to respond to two questions. First, what technical issues must be addressed

in order to accomodate full and public disclosure? Second, what mechanism can test professionals use to communicate effectively with our publics about basic and complex measurement practices?

Some of the technical issues that arise from the requirement for full and public disclosure include equating, test administration, scoring and score reporting. These are the areas that will be most directly affected by full and public disclosure. As such, there is merit in first reviewing the general issues concerned and then raising questions that confront test developers as well as the questions that generally arise from the regulation of testing and related educational practice.

Equating

If one form of a test could be administered repeatedly without giving some examinees the undue advantage over other examinees of having seen it before, test equating would be unnecessary. The need to equate tests is a consequence of the need to have multiple versions of forms of the "same" test.

There are numerous reasons why multiple forms are necessary. In admissions testing, equal opportunity of access may require limited exposure of test forms. Candidates who have previously studied the exact test that they then take "for real" are in a very different position from those who have not previously studied the test. While we do not know the exact quantitative effect that exposure to a test form has on any individual's subsequent performance on that test form over different amounts of time and various repetitions, it is reasonable to assume that practice on a particular set of test items will increase an individual's score on the next administration.

With full and public disclosure mandated, when an individual takes a test, that test, answer key, and answer sheet are then made available to that individual upon request. Once released to the public, that particular test is no longer secure. A test form that has been made available to even

one indivudual can no longer be administered in a testing program where security is required.

If multiple versions of the same test are needed to ensure that, at any one administration, each examinee will see the test for the first time and thus bring to the test administration no direct practice advantage, a method of making comparable the test scores from multiple versions of a test is necessary. Equating is the process that allows us to take scores from one form of a test and compare them to scores obtained from another form of the test. The question facing test professionals then becomes not whether to equate but how to equate.

Attempts to equate tests are complicated by an obvious and very fundamental property of tests. That is, even if test construction practices are designed to create equivalent sets of items for different forms of a test, the intact forms may not elicit equivalent responses from groups of examinees. These differences in group performance may indicate that either the groups were different to start with and the test forms just reveal that difference, or the groups' performances may be the same, but the test forms create differences because the forms themselves differ somewhat. And, without careful study, it is difficult to identify the "cause" of these differences in performance. For example, if I take Test A and you take Test B, and if I get 10 questions right while you get 12 questions right, there is no way to decide on the basis of this information alone whether your skills are better developed than mine or whether Test B is easier than Test A.

Test professionals must always anticpiate that even though test forms are constructed to be parallel (i.e. equivalent in difficulty and content), there is always the possibllity that the forms may not behave equivalently when administered either to the same or to equivalent groups. The reality of this potential variation in the difficulty of different test froms requires that test equating methods be used. Equating is simply a process

for adjusting differences in test difficulty so that scores obtained on different test forms may be directly and meaningfully compared. There are several common methods of test equating used in large-scale testing programs: spiraling, anchor-test equating, pre-section equating, and item-response theory equating.

Spiraling requires that both a secure equated form -- think of it as the old standard -- and a new form to be equated are administered simultaneously to one group of examinees. For example, if one group of 2000 examinees sat for the Weber Test of Applied Wiggetry, half (1000) of the examinees would take the old standard secure form A. The other half would take the new form B. The assignment of test forms to examinees would be made essentially at random.

What does "random" mean in this case? It means that if one randomly samples members from a specified population, the behavior of the sample should be appropriately representative of how the entire population would behave. Similarly, if one wanted to compare the behavior of a group that had had a particular educational experience or treatment with the behavior of a group similar in every other respect except that it had not been given the treatment, random selection of some of the population into the treatment group and some into the control group would allow such comparisons. Thus, random separation of a group of individuals into two or more treatments allows one to attribute any differences in behavior to differences in the treatment. The parallel for "treatments" in this illustration is the test forms in spiral equating.

In an equating study, then, a group of examinees would be separated into "treatment" groups essentially at random. Thus, one treatment group would be administered one form (that is, the old form) of a test; another treatment group would be administered a second form (that is, the new form).

In Figure 1 are eighteen of these 2000 examinees scheduled to take

Insert Figure 1 about here

the Weber Test of Applied Wiggetry. These students were seated and tests (Forms A and B) were distributed at random. If the test forms are of equal difficulty, there should be no differences in average test performance between the two groups. However, if there is a difference, the test forms may be contributing to that difference and adjustments need to be made so that the distribution of scores from the new form match that of the old form, thus rendering the scores reported from the new form comparable to those reported from the old one.

Anchor-test equating requires that all forms of a test have embedded within them a common set of items -- an anchor -- that will not count toward the total score. This anchor set of items, administered to different groups of examinees, is used to measure the differences in abilities of these groups. The information derived in this manner can then be used to estimate how a group would perform on a test it had not taken.

Spiral equating and anchor-test equating are easy to conceptualize and to distinguish if one remembers that spiraling assumes that the different examinee groups are from one population. Thus, any differences among groups can be accounted for by differences in the test forms. Anchor-test equating assumes that examinees represent different populations; the common feature is the anchor set of items. Any differences in performance on the anchor items indicates systematic differences among the populations; these systematic differences can be used to adjust scores on a related test.

To illustrate anchor-test equating, consider two classes of 9th grade students taking the same algebra subtest: subtest A. Subtest A is one part of a longer algebra test that also includes similar subtests B and C. Suppose we want to know how both classes would do on subtests B and C but do not want to administer all three parts, A, B, and C, to both classes. We

can administer subtest B to one class and subtest C to the other. Having already attained both classes' scores on subtest A, we can compare their performances and use that information to predict how each class would have done on the subtest they did not take.

As shown in Figure 2, each examinee takes one subtest used for scoring

Insert Figure 2 about here

and one subtest used for equating. However, unlike Figure 2, the subtests are not labeled on the test. Indeed, the examinee does not know which subtest is used for which purpose.

Another application of anchor-test equating is to adjust for any differences over time in test difficulty, when new forms are developed during that period, differences that may have been introduced by either a change in item writers or a change in the subject matter. In this situation, if a subsection of the test, an anchor -- never used for score reporting -- were administered each year to new examinees, their performance could be compared to the performance of prior examinees. This information could be then used to adjust the scores from the new part of the test used for scoring.

A variation of anchor-test equating, section pre-equating requires that intact sections of a new form (usually one or two at a time) be administered as part of an existing test but not contribute to the score. Unlike anchor-test equating, where the anchor is essentially a miniature form of the entire test, section pre-equating does not require that the intact section be representative of the entire test; it need only be related.

A fourth method of equating, Item-Response Theory (IRT), establishes another set of requirements. First, a group of test items must be similar in content and difficulty. Second, for all intents and purposes the test must be unspeeded (that is, most people should have at least the opportunity to attempt every item). Third, the test items must be independent (that is,

an individual should be able to respond correctly to any item without that response being dependent upon another correct response). And fourth, the test must be at least 25 items long.

Since IRT equating takes data from individual items rather than from total test scores, it may be more flexible, albeit more demanding technically, than either spiraling or anchor-test equating. IRT enables items to be pretested, calibrated (that is, placed on an item scale), and then added to an item pool. A new form would be assembled from this calibrated item pool, thus permitting the new form to be pre-equated.

Disclosure provisions of legislation designed to regulate testing will render all test items that contribute to examinees' scores unsecure. And, the lack of secure reusable sets of items will threaten test users' abilities to draw reasonable comparisons among examinees taking different forms at different times. The issue to be determined, therefore, is the most appropriate equating method for a given testing program. Regardless of the selected method, however, a number of questions arise in the equating process.

- One issue of general concern is the availability of large test administrations suitable for any type of equating study.
- Will candidates be able to detect equating sections of a test? And, if so, will their performance change or differ in the equating section?
- Should all sections have the same timing requirement to help disguise the location of equating sections? If so, what will/ be the impact on equating methods based on the use of old secure forms with different time requirements?

- How many new sections can be administered at each test administration?
- What are the effects within a testing program when equating methods are changed? For example, when a program that has used spiral equating is faced with no more old secure forms, will "drift" occur in the change to anchor or IRT equating?
- And, finally, how can equating methods be explained to test takers and users?

Test Administration

If we assume that test equating facilitates open testing for all tests, and if we assume that the more forms developed of any test, the more costly each test administration will be, then we must ask how test developers can meet the requirements of full and public disclosure and the needs of examinees by not making tests prohibitively expensive.

Minimum test disclosure is desirable to minimize the cost passed on to each examinee. This is a generally worthwhile goal in days of galloping inflation regardless of the product or service discussed. It is, however, particularly critical that the cost of admission tests be kept at a minimum so that economically disadvantaged individuals are not adversely affected. After all, standardized objective instruments have been deliberately used in the admissions process in order to equalize opportunity. Thus, test developers must strive to meet the disclosure requirements without threatening economic discrimination. This is a difficult undertaking which raises such questions as:

1) How many secure test forms already exist?

2) What will it cost to develop each new form in terms of time, staff, materials, and equipment?

3) What further adjustments will be required to develop special forms for handicapped students (for example, the blind)?

4) What further adjustments will be required for special administrations (for example, the Sabbath observer)?

Until these questions are answered for each testing program, decisions about test administrations can not be made, but we can speculate about the impact of some decisions.

For example, what are the ramifications of a decision to reduce test administrations by, for example, 20%? One might be that testing sites for remaining administrations would have to be increased. Another might be that sites would have to be selected that could accommodate large numbers of examinees, as well as special materials, equipment and facilities for handicapped students.

And, what are the ramifications of a decision to reduce only special administrations by 20%, for example? One might argue that since Sabbath observers represent only a very small portion of the examinee population, a reduction in number of test administrations would have the least impact here. But what about the ethical considerations? To reduce special administrations disproportionately for Sabbath observers certainly invites charges of discrimination, parallels hold for other population subgroups as well. Moreover, if the number of test administrations is reduced for any reason, accessibility of each administration will become more important; potential conflicts with major evelts in a given area will have to be considered.

Scoring

Test developers are, in part, responsible for effectively communicating with the publics affected by tests about traditional and innovative test development practices. And by "test development" is meant the entire process -- writing specifications, designing and piloting forms, validating, and disseminating. One part of the process, however, has been conspicuously characterized by a lack of effective communication -- score conversions. Whether trying to explain norms tables, scale scores or any other derived score, we have succeeded only in conveying to the publics served by standardized tests an illusion of mustical manipulation. Examinees may wonder why the raw score, or "number right" is not reported directly. Why is that raw score translated into a scale score, or percentile, or other derived score? A simple answer is that raw scores are seldom directly meaningful.

How can one tell whether "95" is a good score? One would first have to know "95 of how many?" hen either normative information ("how did others in the peer group do?") or criterion-referenced information ("does 95 represent mastery of a skill?") would have to be provided. Derived scores like scale scores and percentiles allow test-score users to interpret score information without knowing specific details about either the number of test items or the observed test characteristics of the peer group.

The lack of effective communication in this area is exacerbated by the ongoing debate between Rights - and Formula - scoring advocates, which has been described by William Angoff:

> The controversy over the preferred mode of scoring has continued without much loss of force for more than fifty years, and articles appear in the literature on this and related subjects with no less frequency today than in earlier years. Names prominent in the psychometric

> field have been arrayed on both sides of the dispute. The considerations persuading the writers on this topic to adopt one position or the other have ranged widely from issues having to do with precision of measurement through the practicalities of administering a testing program under conditions understandable and acceptable to the public, to issues of special tactics, ethics, and morality in the making of tests, differences in personality and their effect on scores on cognitive tests, and considerations of equity and fairness to all examinees. (Angoff, 1980).

Rights scoring simply means that an examinee is given a raw score that is a simple count of the number of items correctly answered. Since no penalty is applied for incorrect responses to items, guessing is implicitly encouraged. In contrast to Rights scoring, Formula scoring does include a penalty for incorrect answers. A typical "Formula" is found in Angoff's paper (op cit).

$$\text{Formula Score} = \text{Number of Right Responses} - \frac{\text{Number of Wrong Responses}}{\text{Number of Choices per Item} - 1}$$

Let it be clearly understood that neither type of scoring is intended to be punitive. Rights scoring reflects all correct responses whether "guessed" or "known", wheras Formula scoring is intended to yield a net score of zero for those examinees who guess randomly on every test item. For example, let us assume that a student guesses at random on an 80-item four-choice test. On the average, that student will "guess" correctly on one-fourth of 80 or 20 items. The formula score reported for that student would be calculated by substituting this information into the given formula and the formula score would be zero.[1] If, on the other hand, that

---

[1] $FS = 20 - \frac{60}{3} = 0$

student does not guess or attempt to answer any items that he or she does not know, the formula score would be twenty[2] since there would be no incorrect responses, only correct responses and omitted responses. Test directions should describe how the test is to be scored, so that the examinee knows whether guessing results in a penalty.

Advocates of Formula scoring may argue that Rights-scored tests are less reliable than Formula-scored ones because there is no penalty for incorrect responses. Reliability is another term for consistency, and there is no reason to expect that item responses that result from random guessing should be consistent.

The differences in scores reported by Rights and Formula scoring are likely to be equivalent in examinees who do not guess. On the other hand, Rights scoring will yield, on the average, higher scores than will Formula scoring if the examinee guesses.

Advocates of Rights scoring argue that their method is easier for the examinee to understand and to implement. Furthermore, if answer sheets and test item keys are to be made available to examinees, they may find it much easier to check the score reported under a Rights system of scoring. Rights scoring advocates also suggest that Formula scoring introduces anxiety since examinees seldom understand the Formula scoring directions read at a test administration.

---

[2] $FS = 20 - \frac{0}{3} = 20$

Another phenomenon associated with Formula scoring is negative scores. If Student A correctly answers only 10 out of 80 test items, then according to the formula:

$$FS = 10 - \frac{70}{3} = -13.33$$

Individuals may therefore earn negative scores in the Formula scoring system.

One might conclude from this discussion that Rights scoring is more disirable than Formula scoring, but the complexity of the issue hinges on the potential impact of any change in the scoring system on the integrity or stability of the measuring scale over forms. In other words, how will a change in the scoring system affect an equating system that has been successfully used over several years? And, can old forms, initially administered under different scoring procedures, be equitably administered under new directions? To the extent that Rights scoring encourages guessing by not penalizing incorrect responses, and Formula scoring discourages guessers, it seems clear that applying Rights directions to an old Formula form would result in higher average scores -- that is, positive scale drift. On the other hand, different scoring practices may cause examinees to adopt different test-taking strategies. In this case, how much and what kind of advance publicity should be planned to help examinees adjust their test-taking behavior? Or, across different testing programs, what are the consequences of one testing program using one type of scoring procedure and another program a different type of scoring procedure?

Reporting to Examinees

The legislative requirement that test producers release to examinees their answer sheets, test items, and key for all items that contribute to the total test score offers test professionals an excellent opportunity to educate the public about measurement. After all, of the millions of examinees

who take admissions tests, for example, how many read technical manuals, attend measurement conferences, and the like? Here, then, is an opportuntiy to introduce basic measurement concepts to an audience with vested and personal interests. The challenge we face is how to explain these concepts simply and accurately. The score-reporting process can serve as one of the most direct and immediate vehicles for this massive and necessary educational effort, an effort that can be enhanced if we anticipate the kinds of questions an examinee might ask. For example:

1) Is this score correct? Were scoring errors made?
2) What does the score mean about my skills?
3) What does the score mean about my chances of being admitted into _____ University?
4) What is the appropriate use of this test score by admissions officers?
5) What are some questions I might ask of users of my score to be sure that it's used appropriately?
6) Is my performance likely to increase if I sit for the exam again? Is my performance likely to increase if I study?
7) What does this test have to do with how well I will do on the job or in school?

An Alternative to Test Scores for Selection

This presents some of the technical consequences of legislation to regulate testing. Surely test developers are skilled enough to accomodate the demands of the legislation and continue to produce high quality tests. However, for the purpose of this paper, I have assumed that some institutions will perceive one of the following:

- attempts to classify students based on scores is inherently evil.
- scores are fallible and should not be used to make decisions about individuals
- available procedures do not measure valid characteristics for the selection process

The latter hypothetical perception is one that test users face in a variety of settings. This is the issue of establishing the validity of the selection model and the validity of the test or tests used as indicators in that model. I call this the "Validity Quest."

The second perception is a recognition of error of measurement and serves as a good reminder that no measurement of human characteristics is exact. I call this the "Fallibility Threat."

The first perception would find us all arguing zealously as one does in a religious debate. And, finally, I call this the "Zealot Menace.

As I don't give much time to reasoning with zealots, I suggest that the Zealot Menace be put aside for less secular discussions. The Fallibility Threat is real and deserving of immediate discussion but it is a discussion appropriate for all types and methods of observations.

Similarly, the Validity Threat applies regardless of the characteristics of the observation methodology. Thus, I would ask you to keep these in mind as I review some alternatives to tests for selection of students into postsecondary programs.

First, let me remind you that the use of test scores in the selection and admissions of students originated as an efficient and accurate estimate of characteristics and achievements thought and shown to be important influences in post-secondary achievement. As such, test scores combined

with other information such as recommendations, grade point average, and the like, assisted in the selection and admissions process.

If it is no longer reasonable to request and use test score information from applicants, what objective information is available to admissions committees? As Cameron Fincher reports (1979, 3):

> "There is little doubt that testing has been a valuable part of educational efforts to discover and develop talent...
>
> But abuses of standardized testing has continued. The professional and beneficent use of test results cannot be guaranteed. Technical competence in test research and development has outpaced professional applications and socially responsible users. Users and consumers have not always been mature and wise in their administration, analysis, and interpretation of tests...
>
> The typical school or college is without an admirable degree of sophistication concerning test technology and uses."

One substitution for nationally known tests such as the American College Test Program and the Scholastic Aptitude Test is that each institution develop and administer its own selection tests. These "tailored" tests would, in some states be subject to truth-in-testing legislation but the relative problems of security for test forms would be reduced simply because the test would be administered to fewer people and, presumably, in fewer locations. And, as the college hopefuls would take more than one college-specific entrance exam, it is also unlikely that applicants would remember items from specific tests.

Of course, separate tests for colleges would place a tremendous burden on the potential applicant. First, the applicant would have to select colleges of interest, then, the applicant would have to register for and to a testing site.

From a technical point of view, both the Validity Quest and the Fallibility Menace would fall with tremendous velocity on admissions offices and support personnel. Thus, we may see mini personalized, tailored SATs erupting from the admissions offices of N colleges and universities to alleviate the controversies now rampant about the use of national standardized tests for selection and admission.

This alternate development process for selection tests may meet the needs of some institutions. However, it is an alternative strapped to the very weaknesses articulated by critiques of standardized national selection tests. What then, are alternatives to "tests" for selection and admissions?

One alternative is a standardized structured screening procedure conducted by trained examiners. Variables to examine may include academic performance, civic experiences, personal characteristics such as "motivation", "goals", "sense of responsibility", and the like.

The development of this standardized structured screening would require the identification of variables important to the success or failure of individuals in the program. Next, the non-test mechanism for quantifying an observation of the variables must be selected and established as appropriate, relevant, and efficient.

The relative importance of these variables would then have to be translated into a point or value system for construction of a composite "score" for cross individual comparisons. The weighing of these variables may be used to exaggerate the importance of non-intellectual measures while minimizing such cognitive variables as previous achievement, grade point average, and the like.

A variety of talents, behaviors, and skills can be demonstrated during interviews if the interviewer is adept. Academic behavior may be evaluated by examining specimen sets submitted by the applicant of recent and documentable academic work.

Thus, a periormance-based screening or selection prodedure has the potential for personalizing the selection information-gathering procedure under standardized and relatively objective circumstances. On the other hand, if the construction of a performanced-based procedure which does not suffer from the Fallibility Menace or the Validity Threat is guaranteed to be expensive and subject to cries of discrimination. And interviews are frequently subject to influences of race, sex, age, and the like.

One application of a performance-based selection program is reported by McKee, Harris, and Swanson, (1979). This system, used to select student for a counselor education in graduate program. This procedure is described as an "effort to integrate traditional criteria and non-intellective criteria within a performance-based model" (p. 73). Included are evalutaition of academic performance, work experience, and personal communication skills. It is upon a composite score of 80 or more out of 100 that an individual is admitted to the counselor education program. Additional information is obtained before admissions decisions are made for students earning scores between 70 and 80.

The relative weights assigned are as follows:

Counselor Education Program Screening - Applicant Evaluation Guidelines

CATEGORY I: Academic Performance and Work Experience	50 points

1. Undergraduate G.P.A. (and GRE) — 10 points
   - 3.5-4.00 = 10 points
   - 3.0-3.49 = 9 points
   - 2.5-2.99 = 8 points
   - 2.0-2.49 = 5 points

   (GRE is a requirement of the Graduate School, which establishes cut-off scores.)

2. Course work and G.P.A. in behavioral sciences — 10 points
   a. For each course, award one point up to a total of 5 points
   b. For G.P.A. award points as follows:
      - 3.5-4.00 = 5 points
      - 3.0-3.49 = 4 points
      - 2.5-2.99 = 3 points
      - Below 2.5 0 points

3. Work experience, volunteer service, graduate work — 15 points

   (Give 5 points for each year in the above. Ask for documentation if not in file.)

4. References (three) — 15 points
   - 1.0 -1.25 = 15 points
   - 1.26-1.75 = 13 points
   - 1.76-2.25 = 11 points
   - 2.26-2.50 = 9 points
   - 2.51-3.00 = 5 points
   - Above 3.00 = 0 points

CATEGORY II. Personal Communication Skills	50 points

1. Essay on Career Goal Expectations (200 words or less) — 15 points
   a. Clarity in stating short-term and long-term goals -10 points
   b. Quality of writing - 5 points

2. Personality test (OP)[3] — 10 points

   On each of the following ten scales, award one (1) point for being within (+-) one standard deviation of the mean and one-half (1/2) point for being within (+-) two standard deviations of the mean.

[3]Notice that test scores do appear among these variables (see Personality Test (OPI).)

SCALE	MEAN	SD
TI	32	5
ES	17	5
CO	18	5
AU	38	3
RO	19	4
IE	30	7
PI	44	8
AL	16	4
AM	26	4
PO	7	3

3. Interview 25 points

Points are awarded individually by two faculty and one student in each of the following five (5) categories:

Understanding of counseling	= 5
Understanding of goals	= 5
Awareness of strengths	= 5
Awareness of weaknesses	= 5
Over-all assessment of candidate	= 5

This type of weighting procedure is similar in concept to the weights assigned variables in typical regression formulas used in student selection and admission. However, in the latter case, the weights are derived empirically as a result of the amount of variance in the dependent variable that each independent variable explains. In the counselor education and similar examples, the weights are assigned judgmentally. This distinction is sufficient to remind us of the Validity Threat to the weights assigned which may or may not be valid.

The authors claim that this performance-based screening procedure has tended to eliminate discrimination against minorities which existed when academic-intellectual performance alone was the criterion for admission.

They further suggest that this procedure has helped identify applicants clarify career goals and has helped faculty get to know those students admitted sooner.

What is left unanswered by McKee, Harris and Swanson (1979) is:

(1) Whether the procedure would work with large numbers of applicants (i.e., more than 250).

(2) The cost of this procedure.

(3) The relationship between the selection variables and successful program completion.

Both test and performance based selection procedures assumes that the number of applicants exceed the number of available spaces in any program. This may or may not be the case.

If we review the enrollment figures from recent years as well as predicted enrollment figures for the decade of the eighties, it is apparent that most admissions problems will not be selection but recruitment. Indeed, many admissions officers will admit every applicant. This phenomenon will cause yet another change in admissions requirements. It will become more important, more useful, more sensible to collect information about applicants which will yield diagnostic and prescriptive data. Traditionally, selective admissions practices are not needed in general as they were during periods of limited opportunity.

For the survival of institutions, information must be collected which will appropriately match a student with a curriculum. This information may come from tests -- removed from the requirements of truth-in-testing legislation because this information is for placement not selection -- from interviews, from self-reports, from observations, and so on.

What is really called for here is what Cameron Fincher (1979) calls the constructive use of tests:

> "New rationales for testing are based on the premise that positive, constructive benefits can be gained from systematic, educationally sound testing.... Testing can yield positive benefits if the local concern is shifted from selection per se to placement, advisement, counseling, and program assessment" (Fincher, p. 5, 1979).

Fincher concludes that the "gist of new rationales may be that the value of any testing or assessment procedure should be judged primarily by the extent to which it aids learning or improves instruction" (Fincher, 1979, 5).

Thus, it appears that the alternatives to tests for selection available - such as a composite score consisting of differentially weighted scores from interviews, past performance, expressed motivation and the like -- are subject to the same types of criticism to which tests are subject:

- Attempts to classify students based on scores is inherently evil.
- Scores are fallible and should not be used to make decisions about individuals
- Available procedures do not measure valid characteristics for the selection process

In addition, performance based procedures are a new methodology and, as such, are subject to the growth pains suffered by psychometric procedures during the past seventy-five years. Certainly there is a place for performance-based selection models. Certainly there is a great deal of research needed regarding the use of performance-based selection models. Certainly it is an alternative selection model but it is not free of our three ghosts: Validity Quest, Fallibility Threat, and Zealot Menace.

Any data which is used to assign individuals to scarce resources or scarce slots must be examined for equity, reliability, and validity. Thus, the alternative reviewed here is not a panacea.

Figure 1

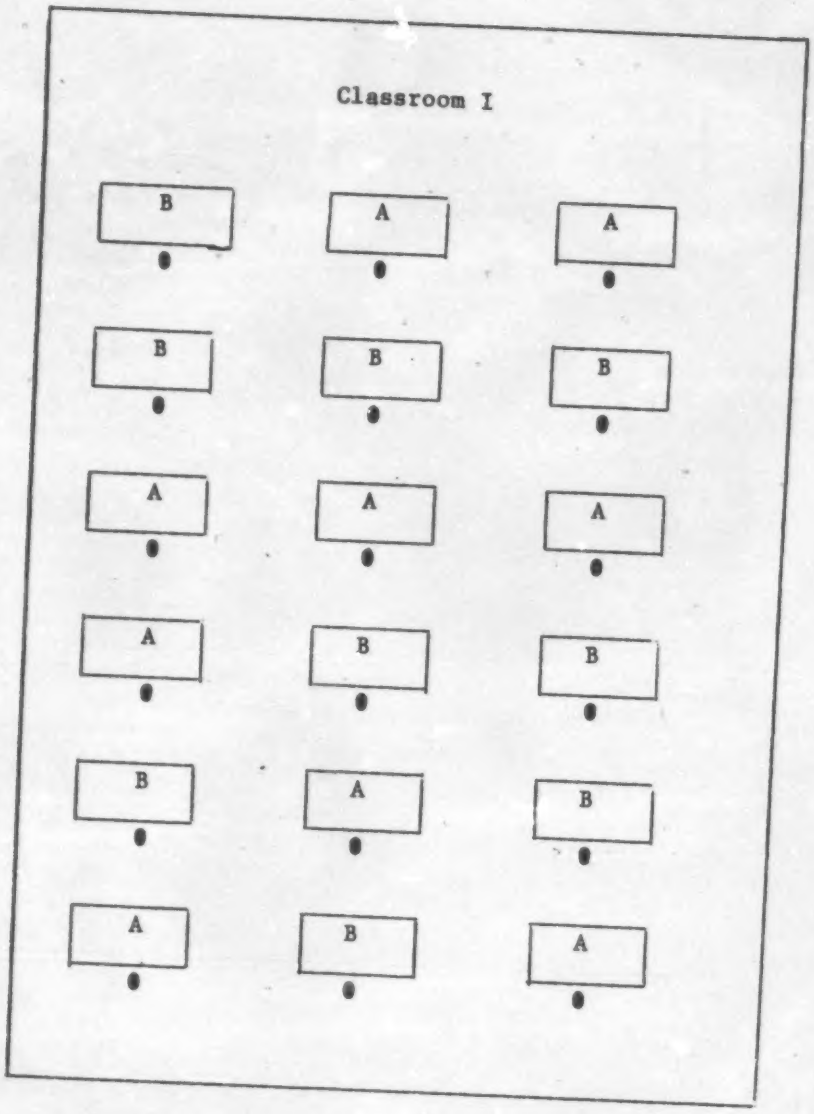
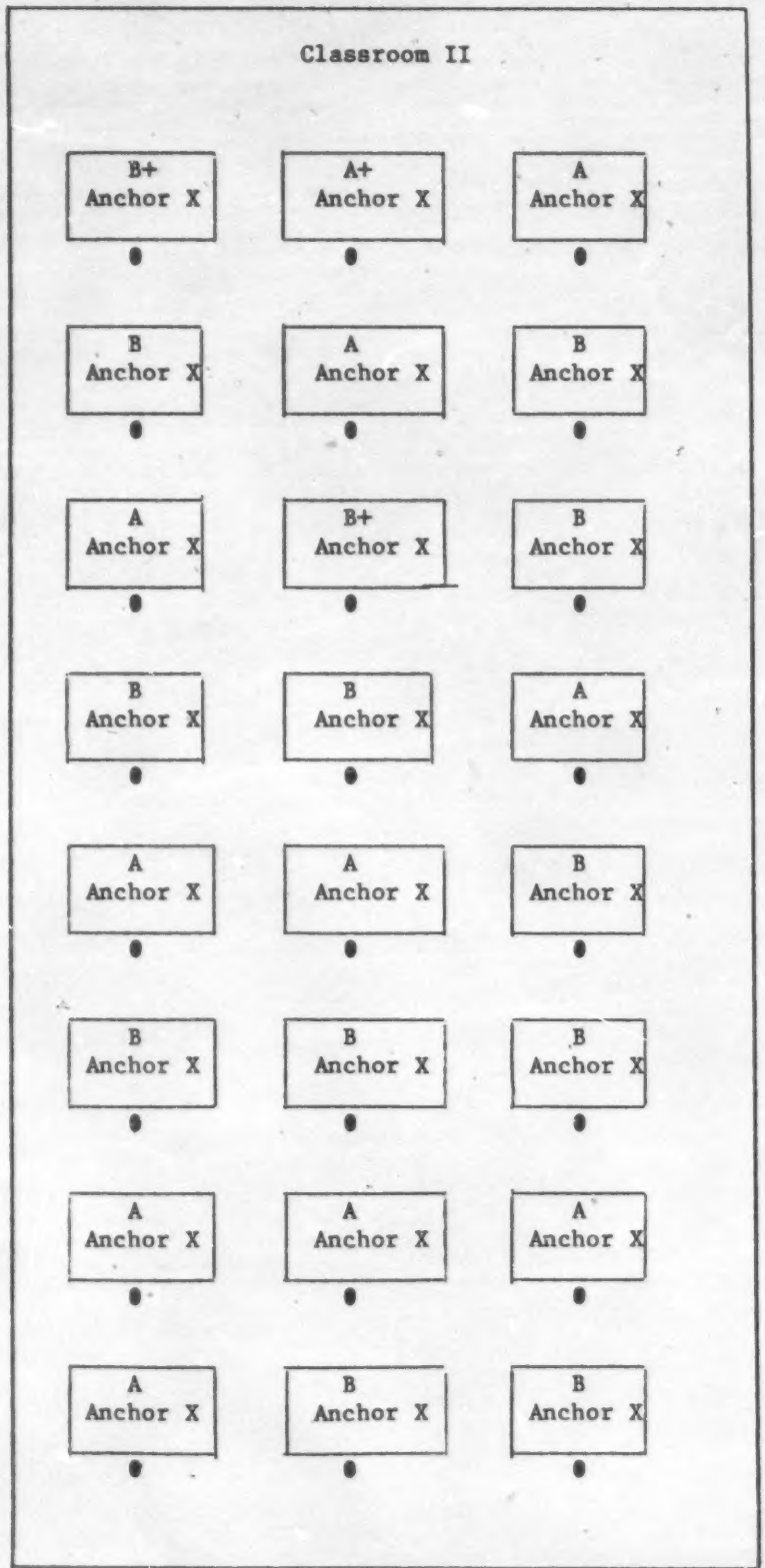


Figure 2


Classroom II
B+ Anchor X
A+ Anchor X
A Anchor X
B Anchor X
A Anchor X
B Anchor X
A Anchor X
B+ Anchor X
B Anchor X
B Anchor X
B Anchor X
A Anchor X
A Anchor X
A Anchor X
B Anchor X
B Anchor X
B Anchor X
B Anchor X
A Anchor X
A Anchor X
A Anchor X
A Anchor X
B Anchor X
B Anchor X

References

Angoff, W. Rights Only Versus Formula Scoring. Educational Testing Service 1980.

Fincher, C. Using Tests Constructively in an Era of Controversy. The College Board Review, Fall 1979, No. 113.

McKee, J. E., Harris, C. M., & Swanson, C. D. A Performance-Based Admissions Screening Program Personnel and Guidance Journal, September 1979.